บทความพิเศษ

Natural Language Processing

ยืน ภู่วรวรรณ

การประมวลภาษาธรรมชาติ (ภาษาไทย)

ศาสตร์ที่รอการพัฒนา

แนวทางในการพัฒนาของภาษาไทย ที่ควบคู่ไปกับพัฒนาการของคอมพิวเตอร์ตามยุคสมัย กับเป้าหมายในอนาคต

การคาดคะเนเทคโนโลยีในอนาคตเป็นเรื่องที่หลายคนอยากรู้ อนาคตคอมพิวเตอร์จะมีความสามารถเพียงไร คอมพิวเตอร์จะเข้ามาช่วยงานอะไรเราได้บ้าง ระบบการทำงานอัตโนมัติต่าง ๆ ที่เราเห็นในภาพยนตร์จะเป็นไปได้เพียงไร

"ในปี ค.ศ. 2021 มนุษย์จะเลิกใช้แป้นพิมพ์อย่างเด็ดขาด คีย์บอร์ดจะเหลือไว้ให้ลูกหลานดูเท่านั้น" การคาดคะเนเทคโนโลยีนี้ดูจะเป็นเรื่องที่ไกลตัวอยู่มากพอควร ถ้าอีกประมาณสามสิบปีไม่มีแป้นพิมพ์ให้ใช้แล้ว เราจะสั่งคอมพิวเตอร์กันอย่างไร สิ่งที่มาทดแทนแป้นพิมพ์จะเป็นอะไร

NLP : ศาสตร์ที่มีผลในอนาคต

"ในช่วงเวลานั้นระบบสั่งงานด้วยคำพูดจะใช้กันอย่างแพร่หลาย คอมพิวเตอร์สามารถเข้าใจภาษามนุษย์ เพียงเราอ่านหนังสือให้เครื่องฟัง เครื่องก็จะพิมพ์ตามเราได้"

นิยายเรื่องนี้เป็นนิยายวิทยาศาสตร์หรือเป็นเรื่องจริงที่จะเกิดขึ้นในเวลานั้น แนวโน้มหลายสิ่งหลายอย่างบ่งบอกเราว่ามีความเป็นไปได้สูง ปัจจุบันเราก็

Natural Language Processing
การประมวลภาษาธรรมชาติ (ภาษาไทย) ศาสตร์ที่รอการพัฒนา

ลดการใช้แป้นพิมพ์ลง หันมาใช้เม้าส์แทน มีการใช้กราฟิกยูสเซอร์อินเตอร์เฟส เทคโนโลยีหันมาทางด้านจินตทัศน์ (Visual) มากขึ้น มีการใช้ออปเจ็กต์ต่าง ๆ บนโปรแกรม การใช้งานได้เปลี่ยนไปจากเดิมเป็นลำดับ

ยิ่งมีการใช้เพนเบสคอมพิวเตอร์กันแล้ว แป้นพิมพ์ก็เลยหายไป กลายเป็นปากกาเขียนตัวอักษรแทน มีการสั่งงานด้วยลายมือเขียน การใช้ปลายปากกานี้จึงเป็นเรื่องที่สะดวกสบายกว่าการใช้แป้นตัวอักษรที่มีจำนวนปุ่มมากมาย แน่นอน อนาคตมีแนวโน้มที่จะเลิกใช้แป้นพิมพ์ค่อนข้างสูง และอาจจะเร็วกว่าที่เราคาดการณ์ไว้เสียอีก

ประมาณปี ค.ศ. 2010 คอมพิวเตอร์มีขีดความสามารถในการแปลภาษาได้คุณภาพงานแปลอยู่ในระดับที่ใช้ในการสื่อสารกันได้ ระบบการแปลภาษาด้วยคอมพิวเตอร์จะมีใช้กันอย่างแพร่หลาย มีโปรแกรมช่วยตกแต่งข้อความ ตรวจสอบการเขียน ผู้ที่ไม่สันทัดในเรื่องภาษาต่างประเทศจะมีซอฟต์แวร์เข้ามาช่วย ทำให้โลกเราเสมือนไร้พรมแดน ทุกชาติทุกภาษามีเครื่องมือที่ช่วยในการสื่อสารให้เข้าใจกันได้ดียิ่งขึ้น

ระบบการแปลภาษาเป็นศาสตร์ที่ต้องใช้เวลา ความเข้าใจภาษาของคอมพิวเตอร์เป็นความใฝ่ฝัน ถึงแม้จะยังไม่เข้าใจได้จนเสมือนการอ่านหนังสือ แต่ก็ใส่ข้อมูลเป็นตัวอักษรให้ก่อน

ประมาณปี ค.ศ. 2000 ระบบมัลติมีเดียเข้ามามีบทบาทมาก มีการเก็บเอกสารข้อความในหนังสือแผ่นซีดีจำนวนมาก คอมพิวเตอร์สามารถประมวลผลภาพ ข้อความ และเสียงในระดับมัลติมีเดียอย่างดี มีระบบการอ่านข้อความที่เป็นเสียงพูดใช้กันอย่างแพร่หลาย เช่น เก็บหนังสือไว้ในซีดี เมื่อเราเรียกค้น คอมพิวเตอร์สามารถอ่านข้อความนั้นออกมาเป็นเสียงพูด พร้อมทั้งแสดงรูปภาพได้อย่างชัดเจน

เทคโนโลยีมัลติมีเดียได้เริ่มต้นขึ้นแล้วในปัจจุบัน เริ่มต้นอย่างช้า ๆ แต่ก็มีความแน่นอน ความจุของหน่วยความจำเริ่มมากขึ้น อิเล็กทรอนิกส์บุ๊ค หรือหนังสือที่เก็บในรูปสื่อ จะมีให้พบเห็นมากขึ้น ในไม่ช้าเราจะเห็นร้านขายหนังสือขายแผ่นซีดีกันมากมาย

ปัจจุบันขีดความสามารถของไมโครคอมพิวเตอร์เริ่มทำงานในสิ่งที่คาดหวังไว้สำหรับอนาคตแล้ว มีการเก็บพจนานุกรมลงในเครื่องอิเล็กทรอนิกส์ขนาดเล็ก มีระบบเวิร์ดโปรเซสเซอร์ที่ชาญฉลาดมากขึ้น การใช้งานคอมพิวเตอร์ใช้หลักการที่ต้องให้ผู้ใช้งานเรียนรู้วิธีการใช้ได้รวดเร็ว และเป็นธรรมชาติมากที่สุด

NLP พัฒนาการที่เริ่มมาพร้อมคอมพิวเตอร์

การประมวลผลภาษาธรรมชาติเป็นสิ่งที่นักพัฒนาระบบคอมพิวเตอร์ได้คิดและพัฒนามาตั้งแต่ครั้งเริ่มต้น ในยุคแรกของการพัฒนา ภาษาที่ใช้เป็นภาษาระดับต่ำ เช่น แอสเซมบลี ต่อมาก็มีการพัฒนาภาษาระดับสูงเช่น ฟอร์แทรน โคบอล ปาสคาล และซี

หากพิจารณาโครงสร้างไวยากรณ์ภาษาที่สร้างขึ้น รูปแบบไวยากรณ์มีลักษณะลงตัวแน่นอน ซึ่งแตกต่างจากภาษาธรรมชาติ การสร้างตัวแปลภาษาจึงมีการพัฒนาอย่างต่อเนื่องกันมา ภาษาที่สร้างขึ้นมีการพัฒนารูปแบบให้ใช้งานง่าย มีความคล่องตัว และสามารถนำมาสร้างระบบงานต่าง ๆ ได้มากมาย

ภาษาคอมพิวเตอร์หรือภาษาที่มนุษย์สร้างขึ้น มีการประมวลผลในระดับต่าง ๆ เช่นเดียวกับภาษาธรรมชาติ มีการประมวลผลในระดับคำ (morphological) ระดับความสัมพันธ์ (syntaclic) และระดับความหมาย (remantic) การประมวลผลของภาษาที่ดำเนินการกันจะอยู่ในกรอบตามกฎเกณฑ์ที่วางไว้

การใช้งานภาษาธรรมชาติอย่างจริงจังก็เริ่มเกิดขึ้นตามการพัฒนาของเทคโนโลยี วิวัฒนาการทางภาษาคอมพิวเตอร์เหมือนกับว่าจะหยุดอยู่ที่การกำหนดภาษามาตรฐาน แต่ความเป็นจริงแล้วการประมวลผลต้องการให้ได้ภาษาที่ใกล้เคียงกับภาษาของมนุษย์ให้มากที่สุด

ยุคเมนเฟรมหรือยุคคอมพิวเตอร์รุ่นแรก ๆ การประมวลผลจึงเป็นไปที่ภาษาสั่งงานคอมพิวเตอร์ เพื่อสร้างรากฐานของการประยุกต์ใช้งานด้านต่าง ๆ ต่อมา

เริ่มยุคแรกจากการทำให้แสดงผลภาษาไทยบนไมโครคอมพิวเตอร์

ในปี พ.ศ. 2521 ผู้เขียนได้มีโอกาสใช้เครื่องไมโครคอมพิวเตอร์ IMSAI รุ่นแรก ๆ ที่ใช้ซีพียู 8080 วิ่งอยู่บน S100 บัส ไมโครคอมพิวเตอร์ขณะนั้นต้องใช้เทอร์มินัลที่เป็นแบบดัมพ์เทอร์มินัล ต่อกับเครื่องพิมพ์ทีทีวาย ความรู้สึกขณะนั้นเห็นว่าไมโครคอมพิวเตอร์เป็นอุปกรณ์ที่จะเข้ามามีบทบาทที่สำคัญแน่นอน เครื่องคอมพิวเตอร์ที่ตั้งอยู่บนโต๊ะทำงานได้ดี และใช้งานได้ง่าย (ขณะนั้นส่วนใหญ่มีภาษาเบสิกและแอสเซมบลีให้ใช้ ยังไม่มีโปรแกรม

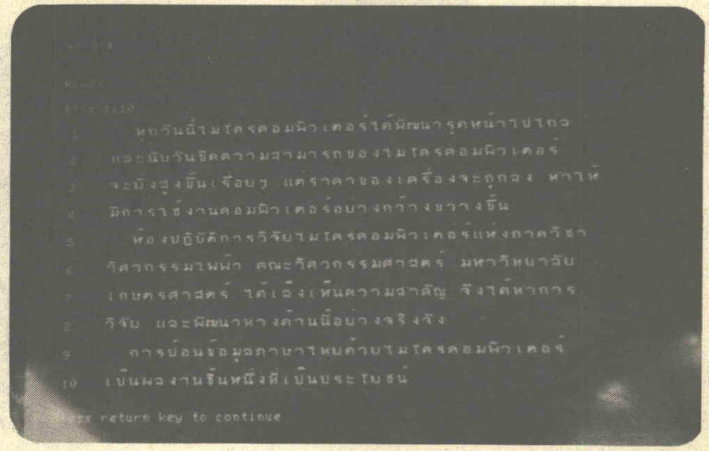
บทความพิเศษ

Natural Language Processing

การประมวลภาษาธรรมชาติ (ภาษาไทย) ศาสตร์ที่รอการพัฒนา

สำเร็จรูป) ความคิดที่จะทำให้เครื่องคอมพิวเตอร์แสดงผลภาษาไทยได้เกิดขึ้นแล้ว

ในปี ค.ศ. 2523 เป็นปีเริ่มต้นที่งานพัฒนาภาษาไทยได้เริ่มต้น ผู้เขียนมีโอกาสไปพัฒนาระบบ Thai text editor ที่ประเทศญี่ปุ่น ในขณะนั้นใช้เครื่อง textronix ซึ่งสามารถสร้างตัวอักขระภาษาไทยในรูปแบบเวกเตอร์ ทำให้การเขียนโปรแกรมเพื่อสั่งงานหรือทำงานภาษาไทยจึงทำได้ง่าย ข้อเสียคือต้องใช้กับเครื่อง Textronic เท่านั้น ไม่สามารถใช้กับเครื่องอื่นได้ การทำงานจัดอยู่ในเกณฑ์ที่ช้า การพิมพ์ตัวอักษรไทยยังอยู่ในรูปแบบเวกเตอร์บนเครื่องพิมพ์พิเศษ Thaitext editor ที่พัฒนาขึ้นในขณะนั้น จึงจัดได้ว่าเป็นเอดิเตอร์ที่ทำขึ้นเป็นต้นแบบเท่านั้น



รูปที่ 1
ไทยเท็กซ์เอดิเตอร์ที่พัฒนาบนเครื่อง Textronic สังเกตตัวอักษรเป็นแบบเวกเตอร์ พัฒนาในปี พ.ศ. 2523

ในปี พ.ศ. 2525 ไมโครคอมพิวเตอร์ได้เข้ามาในประเทศไทยมากแล้ว การพัฒนาภาษาไทยได้ก้าวหน้าขึ้นมาอีกขั้นหนึ่ง การแสดงผลบนเครื่องไมโครคอมพิวเตอร์เริ่มทำได้ดีขึ้น มีการแก้ไขเปลี่ยนแปลงรอม การเปิดตัวอักษรบนเครื่องพิมพ์แบบดอตแมตริกซ์ ทำให้การประมวลผลภาษาไทยได้ดีขึ้น

ปี พ.ศ. 2526 ผู้เขียนและทีมงานได้มีโอกาสเสนอผลงานให้สภาวิจัยแห่งชาติพิจารณาในเรื่องการประมวลผลภาษาธรรมชาติ สิ่งที่เสนอในขณะนั้นเป็นเรื่องเวิร์ดโปรเซสเซอร์ และอัลกอริทึมการจัดวางอักขระบนตารางแอสกี ซึ่งเป็นที่มาของตารางรหัสเกษตร นอกจากนี้ยังได้เสนอวิธีการจัดเรียงคำภาษาไทย ผลงานในปีนั้นได้รับรางวัลจากสภาวิจัย

รูปที่ 2
เวิร์ดโปรเซสเซอร์บนเครื่องออสบอรัน Z80 ไมโครคอมพิวเตอร์ขนาด 8 บิต สมัยแรก

ก้าวเข้าสู่การพัฒนาระดับคำ

เมื่อการประมวลผลภาษาไทยเป็นเรื่องจำเป็นสำหรับคนไทย ในปี พ.ศ. 2526 ผู้เขียนและทีมงานเห็นว่าการศึกษาระดับคำควรเริ่มได้แล้ว จึงได้เริ่มทำการเก็บข้อมูลระดับคำในรูปพจนานุกรม คำที่เก็บเป็นคำที่ใช้ในชีวิตประจำวัน โดยต้องการดูว่าคำที่ใช้ในชีวิตประจำวันมีเป็นจำนวนเท่าไร ถ้าต้องการให้คอมพิวเตอร์เก็บคำเหล่านี้ไว้ใช้จะต้องใช้จำนวนคำเท่าไร ผลงานวิจัยในขณะนั้นทำให้ทราบว่า คำที่ใช้ในชีวิตประจำวันทั่วไปมีไม่เกินหนึ่งหมื่นคำ ผู้เขียนเลือกคำที่ใช้มากไว้เพียง 5,400 คำ โดยจัดเรียงลำดับตาม degree of usage เช่น คำว่า ที่ มีเปอร์เซ็นต์ใช้ 2.2 เปอร์เซ็นต์

การตัดคำไทยเป็นสิ่งท้าทาย

ในช่วงปี 2527 เป็นปีที่มีการพัฒนาเวิร์ดโปรเซสเซอร์ทำได้ดีมากแล้ว เวิร์ดโปรเซสเซอร์ที่พัฒนาในขณะนั้นส่วนใหญ่ยังมีข้อจำกัดอยู่ การเพิ่มขีดความสามารถในเรื่องการตัดคำไทยจึงเป็นสิ่งที่ต้องศึกษาหาวิธี

ช่วงเวลาเดียวกันนั้นกลุ่มผู้วิจัยทั้งหมดได้พยายามสร้างกฎเกณฑ์ทางภาษาในเรื่องการตัดคำ โดยดูจากสถิติที่เก็บไว้มาวิเคราะห์หากฎเกณฑ์ กฎเกณฑ์ที่สร้างขึ้นมีถึง 73 ข้อ แต่ละกฎมีเงื่อนไขยกเว้นอยู่มากมาย กฎแต่ละข้อส่วนใหญ่จะเป็นกฎที่ใช้ได้

สืบเนื่องจากกฎการแบ่งคำนี้มีมาก และยังไม่สามารถบอกได้ว่าจะครอบคลุมไปได้ดีเพียงใด การทดลองการตัดคำด้วยกฎจึงเริ่มขึ้น

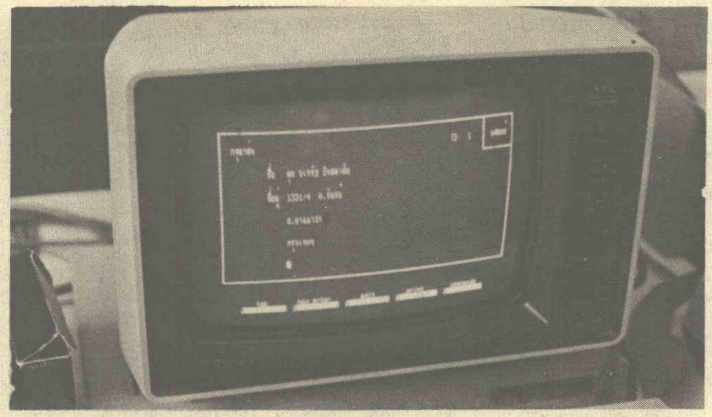
กฎการตัดคำดีแต่ประยุกต์ไม่ได้

เมื่อสร้างกฎไว้มาก ครั้นมาประยุกต์ใช้ปรากฏว่าเขียนโปรแกรมเพื่อตรวจสอบกฎเกณฑ์จำนวนมากไม่ได้ ความเร็วของการประมวลผลไม่เร็วพอ การทดสอบกฎจึงทำได้เพียงในห้องแลป มีการเขียนทดสอบกันด้วยภาษาโปรล็อก กฎต่าง ๆ เหล่านี้จึงยังคงอยู่ในเอกสาร

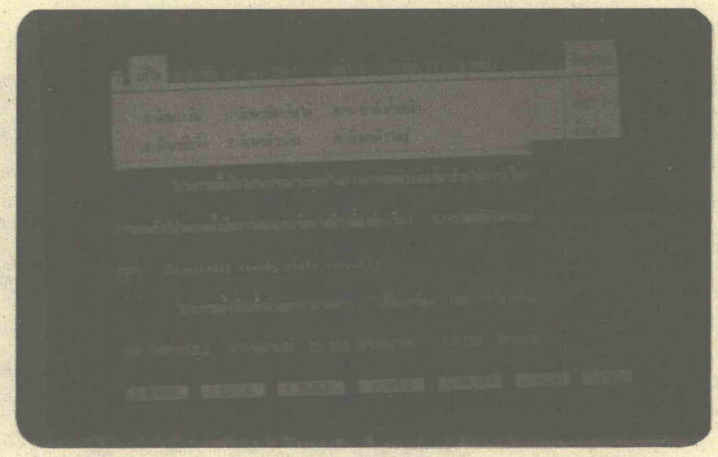
อย่างไรก็ดี มีการแบ่งแยกกฎออกมาประมาณ 15 ข้อ เพื่อนำมาใช้ในเวิร์ดโปรเซสเซอร์ขวัญใจนักพิมพ์ดีดรุ่นแรก เพื่อใช้ในการแบ่งบรรทัด เมื่อกด ctrl B การตัดคำด้วยกฎเพียงเท่านี้จึงยังใช้ไม่ได้ผล มีเปอร์เซ็นต์ถูกต้องเพียงประมาณ 80 เปอร์เซ็นต์ เช่น แบ่งบรรทัด 10 บรรทัด ถูกต้อง 8 บรรทัด อีก 2 บรรทัดยังผิดพลาด และการตัดแบ่งยังเป็นปัญหาอยู่มาก

กฎเกณฑ์ทางภาษาสู้พจนานุกรมไม่ได้

ในปี พ.ศ. 2528 จึงได้มีการทดลองใช้พจนานุกรมในการแบ่งแยกคำ ข้อเสนอของผู้เขียนคือการใช้ longest map with backtraching ซึ่งเป็นอัลกอริทึมที่ใช้ในการแบ่งคำ ผลปรากฏว่าใช้ได้ผลดี มีการนำเอาพจนานุกรมขนาด 5,400 คำ ที่เป็นคำที่ใช้งานในชีวิตประจำวันมาใช้ ผลปรากฏว่าการแบ่งแยกคำได้ผลดีมาก สามารถแก้ไขและกู้คำที่ไม่มีในพจนานุกรมได้ โดยการทำ window scan วิธีการนี้จึงเป็นวิธีการที่ได้ผล และเป็นที่ยอมรับในเวลาต่อมา



รูปที่ 3
พัฒนาการภาษาไทยบนเครื่อง NEC PC8000 รุ่นแรกของการใช้ไมโครคอมพิวเตอร์ 8 บิต



รูปที่ 4
ขวัญใจนักพิมพ์ดีดที่พัฒนาขึ้นบนเครื่อง SiriusI

การทดลองวางรหัสภาษาไทยแบบสองไบต์

หากจะพิจารณาดูแนวโน้มของเทคโนโลยีพบว่า ภาษาไทยจะใช้กับภาษาในค่ายตะวันตก คือ แอสกีอย่างเดียวไม่ได้ เราอาจต้องใช้ร่วมกับภาษาในค่ายตะวันออก คือ ญี่ปุ่นและจีนบ้าง ดังนั้นผู้เขียนจึงเสนอแนวทางการใช้รหัสสองไบต์แทนอักขระภาษาไทย หลักการยึดตามรูปแบบภาษารูปภาพเหมือนภาษาจีนและญี่ปุ่น ในขณะนั้นได้สร้างรูปแบบตัวอักขระที่มองทั้งสี่ระดับ ผลปรากฏว่าสองไบต์ที่ใช้สามารถแทนอักขระไทยได้อย่างมีระบบ มีการศึกษาเปรียบเทียบการประมวลผลในแง่ต่าง ๆ เมื่อเทียบกับรหัสแบบหนึ่งไบต์ เช่น การแสดงผลบนจอภาพ การพิมพ์ผลบนเครื่องพิมพ์ การจัดเรียงคำ การค้นหา การแทรกคำข้อความในรหัส และการประมวลผลในรูปแบบพิเศษ เช่น การเขียนอ่านลงดิสก์ การแปลระหว่างรหัสหนึ่งไบต์กับสองไบต์ ผลงานดังกล่าวเป็นบรรทัดฐานสำหรับอนาคต เผื่อว่าภาษาไทยอาจต้องใช้ร่วมกับภาษาอื่น หรืออาจต้องรวมอยู่ในรหัสเดียวหรือในภาษาโกลบอล

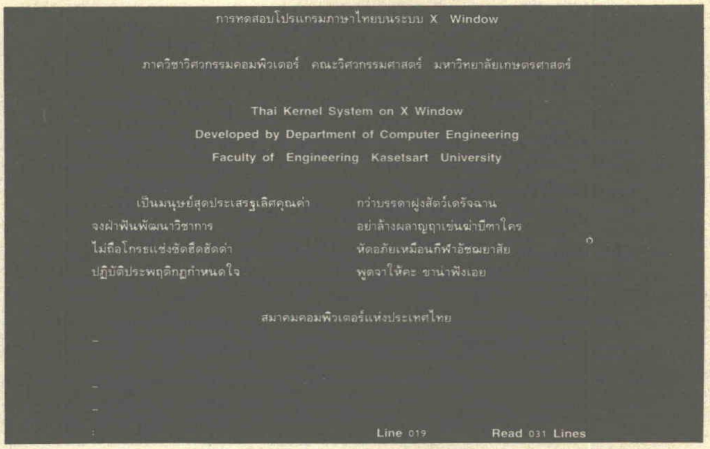
ต้นฉบับไม่มีหน้า 231-232

No This Page 231-232 In Original

การสร้างรากฐานแก่นภาษาไทยในโอเอส

พัฒนาการทางภาษาไทยทำให้เกิดการพัฒนาไทยการ์ดกันอย่างกว้างขวาง มีการพัฒนาแก้ไขฮาร์ดแวร์แสดงผลจนทำให้ระบบไทยการ์ดในประเทศไทยในยุคปี พ.ศ. 2529 เกิดขึ้นอย่างมาก ความสับสนของผู้ใช้ไทยการ์ดเริ่มมีมากขึ้น

ความคิดในการแก้ไขปัญหาดังกล่าวด้านหลักการประมวลผลภาษาไทยจึงเกิดขึ้น ทีมงานวิจัยที่ผู้เขียนมีส่วนเกี่ยวข้อง จึงเสนอโครงการให้เป็นสำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยี (STDB) เพื่อสร้างรากฐานการพัฒนาซอฟต์แวร์ภาษาไทยให้เป็นไปในทิศทางที่ถูกต้อง จึงเสนอให้สร้างระบบภาษาไทยระดับแก่น ที่เรียกว่า TKS เป้าหมายที่สำคัญคือ ต้องการให้เป็นระบบ API ระบบหนึ่งของการพัฒนาโปรแกรมประยุกต์ภาษาไทย TKS-Thai Kevnel System จึงเป็นซอฟต์แวร์ระดับล่าง เพื่อครอบคลุมการประมวลผลภาษาไทยไว้ ผู้เขียนโปรแกรมประยุกต์ไม่ต้องยุ่งยากในการเขียนจัดการภาษาไทย โครงสร้างลักษณะจึงเหมือนกับผู้พัฒนาโปรแกรมประยุกต์บนวินโดวส์ขณะนี้



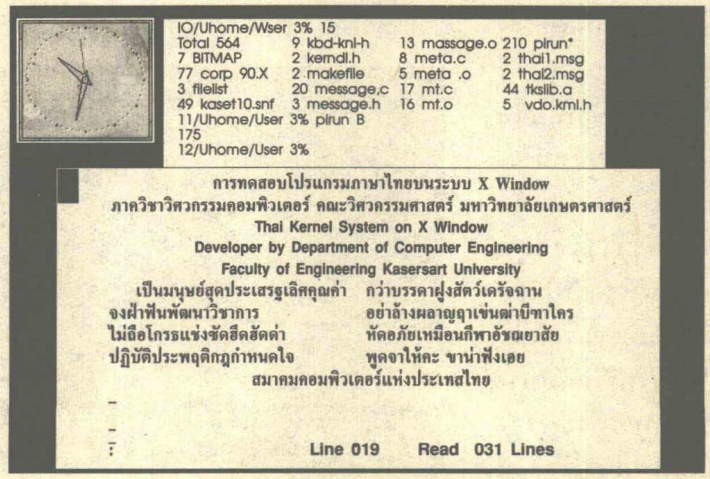
รูปที่ 5 การใช้ TKS บน X วินโดวส์ แสดงการ Protable Application จากดอสไปยูนิกส์ ผ่าน TKS

งานแปลภาษาด้วยคอมพิวเตอร์

ในช่วงปี พ.ศ. 2529-2530 ญี่ปุ่นโดยองค์กร CICC ได้เข้ามาตกลงร่วมมือกับกระทรวงวิทยาศาสตร์ โดยศูนย์อิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติในอันที่จะร่วมมือกันพัฒนาระบบแปลภาษาด้วยคอมพิวเตอร์ ระบบแปลภาษาที่ญี่ปุ่นต้องการคือ แปลได้หลายภาษา ดังนั้นญี่ปุ่นจึงต้องทำวิจัยร่วมกับเจ้าของภาษา เช่น ไทย มาเลย์ อินโดนีเซีย และจีน เป็นต้น

โครงการแปลภาษาในยุคเริ่มต้นเป็นโครงการร่วมมือของมหาวิทยาลัยไทยสี่แห่ง คือ มหาวิทยาลัยเกษตรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าลาดกระบังและธนบุรี โครงการวิจัยดังกล่าวดำเนินไปอย่างเชื่องช้า การแปลภาษาที่เสนอโดยญี่ปุ่นนี้เป็นหลักการที่ญี่ปุ่นเองอยากทดลอง และต้องการเก็บข้อมูลภาษาต่าง ๆ เอาไว้

อย่างไรก็ดี ในฐานะที่ผู้เขียนและทีมงานได้ร่วมอยู่ด้วยในโครงการดังกล่าว แต่ได้รับมอบหมายให้พัฒนาทางด้านอินพุตเอาต์พุต ที่จะแสดงผลภาษาไทย หลังจากนั้นจึงได้ทดลองทำระบบแปลภาษาขนาดเล็กโดยใช้นิสิตทางด้านภาษาศาสตร์และวิศวกรรมศาสตร์ในระดับปริญญาโท มีการทำวิทยานิพนธ์ในหัวข้อดังกล่าวเพื่อศึกษาพื้นฐานโครงสร้างระบบแปล โดยกรอบการแปลอยู่ในขอบเขตจำกัดมาก งานพัฒนาระบบการแปลภาษาเน้นที่ภาษาอังกฤษเป็นไทย เพราะเป็นภาษาที่เราเข้าใจโครงสร้างและไวยากรณ์ได้ดี การทดลองจึงเน้นเฉพาะหลักการทางทฤษฎีเท่านั้น



รูปที่ 6 การแสดงผลภาษาไทยบน X วินโดวส์

งานวิจัยเหล่านี้ยังคงเป็นรากฐานที่ดำเนินการต่อเนื่องได้อีกมาก

งานพจนานุกรมภาษาอังกฤษ-ไทย ในระบบตรวจสอบตัวสะกด

เมื่องานประมวลผลภาษาธรรมชาติในเรื่องที่จะประยุกต์ในยุคแรกได้แก่ ระบบตรวจสอบตัวสะกดในเวิร์ดโปรเซสเซอร์ งานพัฒนาพจนานุกรมจึงเริ่มขึ้นในปี พ.ศ. 2530 ทีมงานวิจัยของผู้เขียนได้เสนอวิธีการตรวจสอบตัวสะกดในการประชุมวิชาการทางวิศวกรรมไฟฟ้าของมหาวิทยาลัยทั่วประเทศ ระบบการตรวจสอบตัวสะกดใช้วิธีการเปรียบเทียบกับพจนานุกรม หากไม่พบในพจนานุกรมให้แสดงเครื่องหมายบอกไว้ ความยุ่งยากจึงเหมือนกับการแบ่งแยกคำไทย การตรวจสอบตัวสะกดจึงต้องการพจนานุกรมที่ดี

หันมาเก็บข้อมูลพจนานุกรมฉบับราชบัณฑิตยสถาน

ในปี พ.ศ. 2531 ศูนย์อิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ กระทรวงวิทยาศาสตร์ ได้ให้ทุนการวิจัยเพื่อจัดเก็บข้อมูลพจนานุกรมฉบับราชบัณฑิตยสถานไว้ในคอมพิวเตอร์ การจัดเก็บข้อมูลนี้มีการวางโครงร่างข้อมูลให้เรียกค้นได้ง่าย ผลงานที่จัดเก็บและการเรียกค้นหาได้อย่างสมบูรณ์ แต่เมื่อทางกระทรวงฯ ได้ปรึกษาหารือกับทางราชบัณฑิต ผลปรากฏว่าทางกระทรวงได้รับการแจ้งว่ายังไม่สมควรที่จะให้เผยแพร่เพราะยังติดขัดด้วยกฎหมายลิขสิทธิ์ ข้อมูลพจนานุกรมนี้จึงเป็นข้อมูลที่ได้รับการนำมาใช้ในห้องแลปต่อไปเท่านั้น

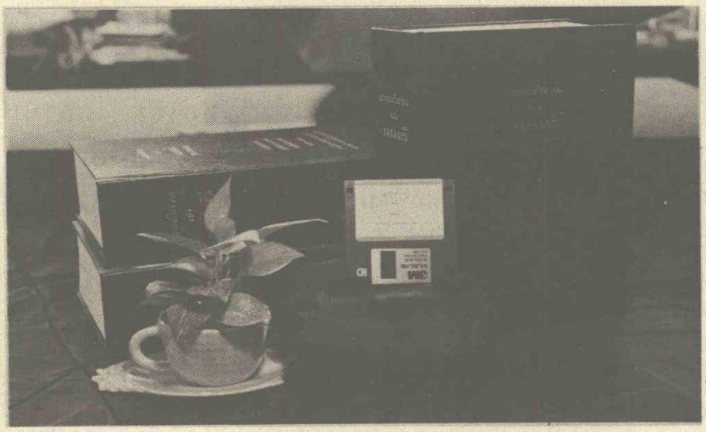
อรรถาภิธาน ฉบับนนทรี (Thai Thesaurus)

งานทางด้านการใช้คำที่มีความหมายเดียวกันในงานเขียนเป็นเรื่องสำคัญ ในภาษาอังกฤษมี Thesaurus ให้ใช้ แต่ในภาษาไทยยังไม่มีใครทำ Thesaurus คณะผู้วิจัยทาง NLP เห็นความสำคัญในเรื่องนี้ โดยเฉพาะหากทำ Thesaurus ที่มีคอนเส็ปต์ทางความหมายร่วมระหว่างภาษาไทย-อังกฤษได้ จะมีประโยชน์ต่องานแปลภาษาด้วยคอมพิวเตอร์ต่อไป

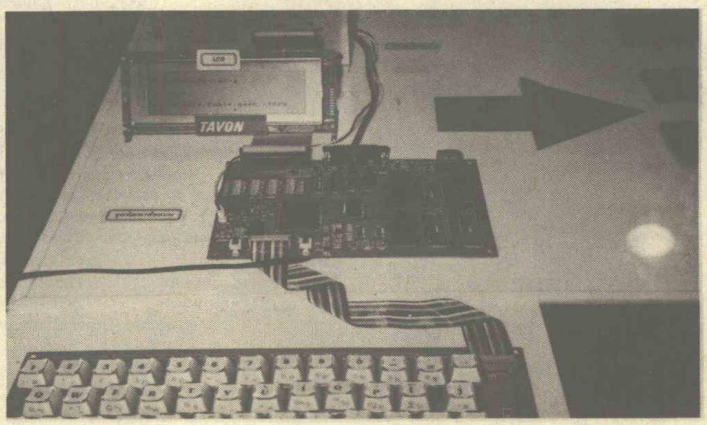
งานนี้จึงเริ่มต้นตั้งแต่ปี พ.ศ. 2531 ใช้เวลาดำเนินการประมาณ 3 ปีก็สำเร็จเป็นอรรถาภิธาน ฉบับนนทรี เล่มแรกขึ้น ปัจจุบันได้จัดพิมพ์เป็นเล่มโดยบริษัท ซีเอ็ดยูเคชั่น จำกัด

การจัดทำ Thesaurus ภาษาไทยนี้ต้องใช้บุคลากรจำนวนมาก คณาจารย์จากภาควิชาภาษาเข้ามาดำเนินการอย่างมาก มีการใช้เทคนิคการรวบรวมเข้าฐานข้อมูล การแก้ไขและเพิ่มเติมปรับปรุงกระทำบนฐานข้อมูล ทำให้งานพัฒนาโครงการนี้ดำเนินการด้วยดี

อรรถาภิธาน ฉบับนนทรีนี้ เป็นการรวบรวมคำศัพท์ไว้มากกว่า 180,000 คำ เมื่อสร้างเป็น main entry ไว้ค้นหา



รูปที่ 7
พจนานุกรมอรรถาภิธาน ฉบับนนทรี หรือ Thesaurus ภาษาไทย เล่มแรก มีคำศัพท์ถึง 180,000 คำ



รูปที่ 8
เครื่องต้นแบบพจนานุกรมอิเล็กทรอนิกส์ขนาดกระเป๋า (ปาล์มท็อป) ที่ผลิตออกจำหน่ายโดยบริษัทฌาวรคอมพิวเตอร์

มีคำมากกว่า 20,000 entry เป็นหนังสือที่รวบรวมคำที่มีมโนภาพเดียวกัน ปัจจุบันอรรถาภิธานมีอยู่ในรูปแบบที่เป็นอิเล็กทรอนิกส์ด้วย

งานประมวลผลภาษาไทยที่ประยุกต์จากอรรถาภิธานยังทำได้อีกมาก ผลงานนี้จะเกี่ยวข้องกับงานด้านการใช้ความหมาย ซึ่งมีความจำเป็นมากต่อการประมวลผลด้วยคอมพิวเตอร์

การพัฒนาพจนานุกรมอังกฤษ-ไทย ฉบับกระเป๋า

เมื่อฐานความรู้ทางด้านพจนานุกรมที่จัดเก็บไว้มีมาก ในปี พ.ศ. 2533 ทีมงานวิจัยจึงร่วมมือกับบริษัทถาวรคอมพิวเตอร์เพื่อพัฒนาพจนานุกรมอิเล็กทรอนิกส์

ขั้นต้นทีมงานวิจัยได้เสนอบริษัทถาวรคอมพิวเตอร์ ให้ผลิตการ์ดพจนานุกรมที่เสียบลงบนสล้อตของพีซี การ์ดต้นแบบได้ดำเนินการเสร็จเรียบร้อยแล้ว ทางบริษัทเห็นว่าการ์ดอิเล็กทรอนิกส์ไม่สะดวก เพราะว่าต้องการพกพา จึงเสนอรูปแบบการพัฒนาเป็นคอมพิวเตอร์แบบมือถือ

งานพัฒนาเริ่มต้นขึ้นทันทีภายในระยะเวลาไม่กี่เดือน เครื่องต้นแบบก็สำเร็จ หลังจากนั้นงานพัฒนาในระดับการผลิตก็เริ่มขึ้น ปัจจุบันบริษัทถาวรคอมพิวเตอร์ได้ผลิตออกขายไปแล้วเป็นจำนวนมาก นั้นเป็นผลงานวิจัยที่เกี่ยวข้องกับ NLP ที่ประสบผลสำเร็จโครงการหนึ่ง

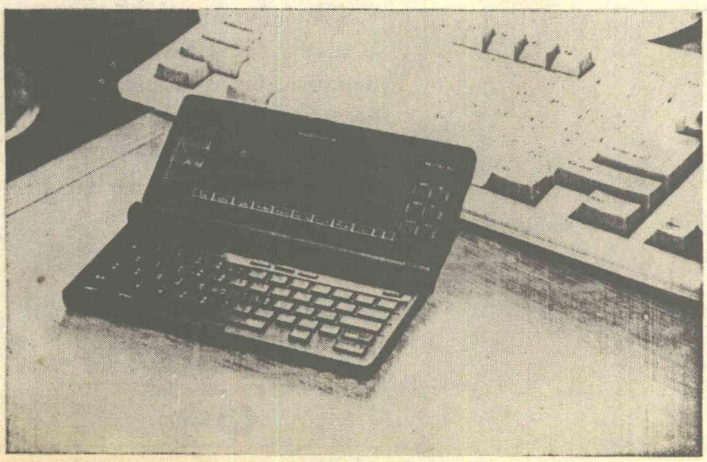
ระบบการตรวจสอบการผิดพร่องการใช้ภาษาไทย

งานประมวลผลภาษาธรรมชาติที่เป็นภาษาไทยยังคงดำเนินต่อไป ในปี พ.ศ. 2535 ดร.อัศนีย์ ก่อตระกูล ร่วมกับคณาจารย์จากภาควิชาภาษาศาสตร์ ได้หากฎเกณฑ์ไวยากรณ์ภาษาไทย และพัฒนาระบบ Parser ภาษาไทย จุดมุ่งหมายของงานคือ การพัฒนาระบบตรวจสอบการผิดพร่องการใช้ภาษาไทย งานวิจัยและพัฒนานี้ยังเน้นในเรื่อง style check เพื่อนำมาใช้ร่วมกับเวิร์ดโปรเซสเซอร์

โต๊ะช่วยงานแปล (translation workbench)

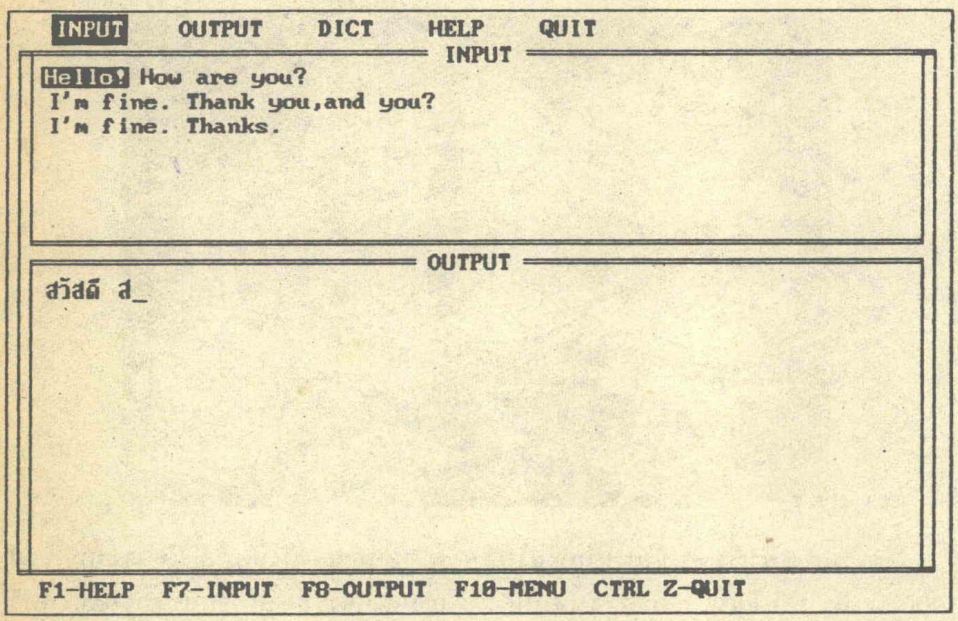
เมื่อเทคโนโลยีงานแปลภาษาด้วยคอมพิวเตอร์ยังต้องการพัฒนาที่หนทางอีกยาวไกล แต่งานพัฒนาบางส่วนอาจนำมาใช้ประโยชน์ได้ คณะผู้วิจัยจึงเสนอแนวทางในเรื่องโต๊ะช่วยงานแปล

ระบบช่วยงานแปลนี้ ใช้คอมพิวเตอร์ช่วย โดยการเก็บพจนานุกรมอังกฤษไทยไว้ ระบบจะเปิดช่องหน้าต่างไว้สามช่อง ช่องหนึ่งเป็นเอกสารต้นทาง ช่องที่สองเป็นช่องที่จะแสดงผลลัพธ์ที่แปล ช่องที่สามเป็นช่องช่วยในการเปิดพจนานุกรม การแปลจะทำงานช่องที่สอง ผู้ใช้ ใช้งานด้วยเอดิเตอร์บนช่อง



รูปที่ 9
ขนาดของพจนานุกรมอังกฤษ-ไทย ไทย-อังกฤษเมื่อเทียบกับแป้นพิมพ์คอมพิวเตอร์

Natural Language Processing
การประมวลภาษาธรรมชาติ (ภาษาไทย) ศาสตร์ที่รอการพัฒนา



รูปที่ 10โต๊ะช่วยการแปลแบ่งเป็น 3 ช่องหน้าต่าง สร้าง Dictionary มีไว้แสดงข้อมูลเกี่ยวกับคำศัพท์

ที่สอง โดยชี้ไปยังคำใดก็ได้บนช่องหนึ่ง จะปรากฏเหมือนเปิดพจนานุกรมบนช่องที่สาม ผู้ใช้เลือกคำจากช่องสามมาใส่ช่องสองได้

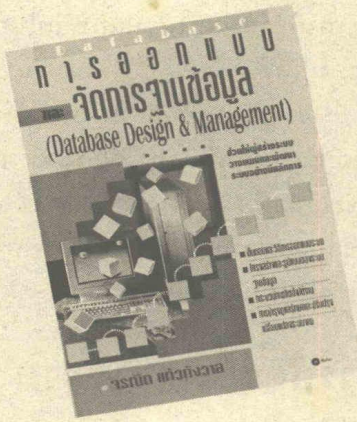
ระบบช่วยงานแปลนี้เป็นซอฟต์แวร์ที่จะใช้ประโยชน์ได้จริง เพราะการแปลยังคงใช้มนุษย์เป็นผู้แปล โดยที่ให้คอมพิวเตอร์ช่วยเปิดพจนานุกรมให้ และมีเอดิเตอร์ช่วยในการเขียนหรืออาจมีเครื่องมืออื่น ๆ ประกอบอยู่บ้าง

หนทางงาน NLP ภาษาไทยยังอีกยาวไกล

ทีมงานนักวิจัยจากห้องปฏิบัติการวิจัยปัญญาประดิษฐ์ยังต้องเดินหน้าต่อไปอีกยาวไกล งานหลายอย่างคงไว้ซึ่งหลักการและทฤษฎีที่จะนำไปใช้ต่อไปในอนาคต งานด้านประมวลผลข้อมูลข่าวสารภาษาไทยจึงยังมีเรื่องที่จะทำอีกมากตราบเท่าที่เรายังใช้ภาษาไทยเป็นภาษาประจำชาติของเรา

งานวิจัยและพัฒนาจะต้องดำเนินการต่อไป ถึงแม้ว่าเป้าหมายการคาดการณ์ทางเทคโนโลยีจะเป็นเช่นไร และอยากเน้นให้เป็นคุณค่าของภาษาไทย ภาษาที่ทำให้เรามีเอกลักษณ์มีความภูมิใจและรวมน้ำใจของคนทั้งชาติไว้

ช่วยให้ผู้สร้างระบบ
วางแผนและพัฒนา
ระบบอย่างมีหลักการ



การออกแบบ
และจัดการฐานข้อมูล
(Database Design & Management)

นำเสนอ ทฤษฎีและหลักการพื้นฐานของระบบฐานข้อมูล ไปพร้อม ๆ กับแสดงตัวอย่างการประยุกต์ใช้งานจริง โดยเฉพาะในด้านการออกแบบและการใช้ข้อมูลทางธุรกิจ โดยไม่ได้มุ่งเน้นที่ตัวทฤษฎีและหลักการของฐานข้อมูลที่ซับซ้อน แต่เป็นการเสริมสร้างความเข้าใจพื้นฐาน และนำเอาทฤษฎีและหลักการต่าง ๆ มาใช้แก้ปัญหาการออกแบบและจัดการฐานข้อมูลในโลกของการทำงานจริง

นอกจากนั้น จะช่วยให้ผู้สร้างระบบ วางแผนและพัฒนาระบบอย่างมีหลักการ มีความเข้าใจในความต้องการของผู้ใช้ระบบในระดับต่าง ๆ กันในองค์การในขณะเดียวกันผู้บริโภคระบบ จะมองเห็นภาพรวมของกระบวนการออกแบบและจัดการระบบ เข้าใจในขีดความสามารถของระบบงาน สามารถเป็นกรรมการหรือมีส่วนร่วมในการสร้างระบบ และใช้ระบบได้อย่างมีประสิทธิภาพ

หาซื้อได้ตามร้านหนังสือชั้นนำทั่วไป
มีจำหน่ายแล้ว !

ซีเอ็ด